Predictors of Customer Satisfaction with Cereal

Shaun Stearns

8/21/2019

# Load Libraries

library(rpart)  
library(rpart.plot)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(adabag)

## Loading required package: foreach

## Loading required package: doParallel

## Loading required package: iterators

## Loading required package: parallel

library(jtools)  
library(sandwich)

# Set Working Directory

setwd("~/Desktop/CSU Global Data Analytics/MIS510/Module 2/")

# Import, view rows, and see list of variables

Cereals.df <- read.csv("Cereals.csv", header = TRUE)  
head(Cereals.df, 9)

## name mfr type calories protein fat sodium fiber  
## 1 100%\_Bran N C 70 4 1 130 10.0  
## 2 100%\_Natural\_Bran Q C 120 3 5 15 2.0  
## 3 All-Bran K C 70 4 1 260 9.0  
## 4 All-Bran\_with\_Extra\_Fiber K C 50 4 0 140 14.0  
## 5 Almond\_Delight R C 110 2 2 200 1.0  
## 6 Apple\_Cinnamon\_Cheerios G C 110 2 2 180 1.5  
## 7 Apple\_Jacks K C 110 2 0 125 1.0  
## 8 Basic\_4 G C 130 3 2 210 2.0  
## 9 Bran\_Chex R C 90 2 1 200 4.0  
## carbo sugars potass vitamins shelf weight cups rating  
## 1 5.0 6 280 25 3 1.00 0.33 68.40297  
## 2 8.0 8 135 0 3 1.00 1.00 33.98368  
## 3 7.0 5 320 25 3 1.00 0.33 59.42551  
## 4 8.0 0 330 25 3 1.00 0.50 93.70491  
## 5 14.0 8 NA 25 3 1.00 0.75 34.38484  
## 6 10.5 10 70 25 1 1.00 0.75 29.50954  
## 7 11.0 14 30 25 2 1.00 1.00 33.17409  
## 8 18.0 8 100 25 3 1.33 0.75 37.03856  
## 9 15.0 6 125 25 1 1.00 0.67 49.12025

t(t(names(Cereals.df)))

## [,1]   
## [1,] "name"   
## [2,] "mfr"   
## [3,] "type"   
## [4,] "calories"  
## [5,] "protein"   
## [6,] "fat"   
## [7,] "sodium"   
## [8,] "fiber"   
## [9,] "carbo"   
## [10,] "sugars"   
## [11,] "potass"   
## [12,] "vitamins"  
## [13,] "shelf"   
## [14,] "weight"   
## [15,] "cups"   
## [16,] "rating"

# Limit dataset to quantitative variables

Cereals2.df <- Cereals.df[c(4:16)]  
head(Cereals2.df, 9)

## calories protein fat sodium fiber carbo sugars potass vitamins shelf  
## 1 70 4 1 130 10.0 5.0 6 280 25 3  
## 2 120 3 5 15 2.0 8.0 8 135 0 3  
## 3 70 4 1 260 9.0 7.0 5 320 25 3  
## 4 50 4 0 140 14.0 8.0 0 330 25 3  
## 5 110 2 2 200 1.0 14.0 8 NA 25 3  
## 6 110 2 2 180 1.5 10.5 10 70 25 1  
## 7 110 2 0 125 1.0 11.0 14 30 25 2  
## 8 130 3 2 210 2.0 18.0 8 100 25 3  
## 9 90 2 1 200 4.0 15.0 6 125 25 1  
## weight cups rating  
## 1 1.00 0.33 68.40297  
## 2 1.00 1.00 33.98368  
## 3 1.00 0.33 59.42551  
## 4 1.00 0.50 93.70491  
## 5 1.00 0.75 34.38484  
## 6 1.00 0.75 29.50954  
## 7 1.00 1.00 33.17409  
## 8 1.33 0.75 37.03856  
## 9 1.00 0.67 49.12025

data.frame(mean=sapply(Cereals2.df, mean, na.rm=TRUE),   
 sd=sapply(Cereals2.df, sd, na.rm=TRUE),   
 min=sapply(Cereals2.df, min, na.rm=TRUE),   
 max=sapply(Cereals2.df, max, na.rm=TRUE),   
 median=sapply(Cereals2.df, median, na.rm=TRUE),   
 length=sapply(Cereals2.df, length),  
 miss.val=sapply(Cereals2.df, function(x)   
 sum(length(which(is.na(x))))))

## mean sd min max median length  
## calories 106.883117 19.4841191 50.00000 160.00000 110.00000 77  
## protein 2.545455 1.0947897 1.00000 6.00000 3.00000 77  
## fat 1.012987 1.0064726 0.00000 5.00000 1.00000 77  
## sodium 159.675325 83.8322952 0.00000 320.00000 180.00000 77  
## fiber 2.151948 2.3833640 0.00000 14.00000 2.00000 77  
## carbo 14.802632 3.9073256 5.00000 23.00000 14.50000 77  
## sugars 7.026316 4.3786564 0.00000 15.00000 7.00000 77  
## potass 98.666667 70.4106360 15.00000 330.00000 90.00000 77  
## vitamins 28.246753 22.3425225 0.00000 100.00000 25.00000 77  
## shelf 2.207792 0.8325241 1.00000 3.00000 2.00000 77  
## weight 1.029610 0.1504768 0.50000 1.50000 1.00000 77  
## cups 0.821039 0.2327161 0.25000 1.50000 0.75000 77  
## rating 42.665705 14.0472887 18.04285 93.70491 40.40021 77  
## miss.val  
## calories 0  
## protein 0  
## fat 0  
## sodium 0  
## fiber 0  
## carbo 1  
## sugars 1  
## potass 2  
## vitamins 0  
## shelf 0  
## weight 0  
## cups 0  
## rating 0

# Regression Tree Predicting Customer Satisfaction

Cereals.ct <- rpart(rating ~ ., data = Cereals2.df, method = "anova",  
 cp = 0.00001, minsplit = 5, xval = 5)  
length(Cereals.ct$frame$var[Cereals.ct$frame$var == "<leaf>"])

## [1] 27

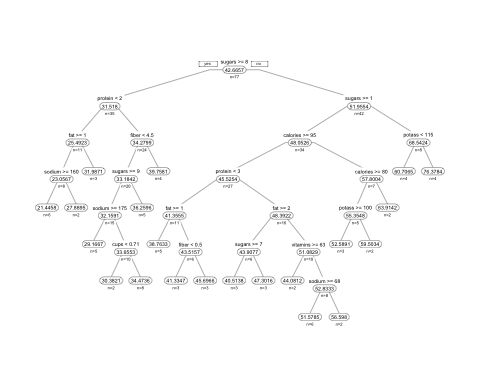
printcp(Cereals.ct)

##   
## Regression tree:  
## rpart(formula = rating ~ ., data = Cereals2.df, method = "anova",   
## cp = 1e-05, minsplit = 5, xval = 5)  
##   
## Variables actually used in tree construction:  
## [1] calories cups fat fiber potass protein sodium sugars   
## [9] vitamins  
##   
## Root node error: 14997/77 = 194.76  
##   
## n= 77   
##   
## CP nsplit rel error xerror xstd  
## 1 0.53171477 0 1.000000 1.03624 0.211948  
## 2 0.18130009 1 0.468285 0.60397 0.139849  
## 3 0.05585013 2 0.286985 0.40080 0.088343  
## 4 0.03883971 3 0.231135 0.41989 0.088549  
## 5 0.03275460 4 0.192295 0.33777 0.076374  
## 6 0.02152225 5 0.159541 0.36524 0.092965  
## 7 0.01287375 6 0.138018 0.34436 0.091881  
## 8 0.01160261 7 0.125145 0.32977 0.079901  
## 9 0.00960558 8 0.113542 0.33224 0.079859  
## 10 0.00817244 9 0.103937 0.32564 0.080166  
## 11 0.00697900 10 0.095764 0.34158 0.083083  
## 12 0.00460845 11 0.088785 0.33130 0.083169  
## 13 0.00434134 12 0.084177 0.34175 0.083455  
## 14 0.00415307 14 0.075494 0.34309 0.084342  
## 15 0.00410716 15 0.071341 0.34309 0.084342  
## 16 0.00382547 16 0.067234 0.33811 0.084053  
## 17 0.00252015 17 0.063408 0.33232 0.075634  
## 18 0.00190303 18 0.060888 0.32577 0.075614  
## 19 0.00178599 19 0.058985 0.32596 0.075452  
## 20 0.00121232 20 0.057199 0.32071 0.075466  
## 21 0.00094050 21 0.055987 0.32608 0.077635  
## 22 0.00065682 22 0.055046 0.32634 0.077622  
## 23 0.00050212 23 0.054389 0.32743 0.077650  
## 24 0.00026003 24 0.053887 0.32819 0.077638  
## 25 0.00021061 25 0.053627 0.32819 0.077638  
## 26 0.00001000 26 0.053417 0.32903 0.077669

pruned.ct <- prune(Cereals.ct,  
 cp = Cereals.ct$cptable[which.min(Cereals.ct$cptable[,"xerror"]), "CP"])  
length(pruned.ct$frame$var[pruned.ct$frame$var == "<leaf>"])

## [1] 21

prp(pruned.ct, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10, digits=-6)

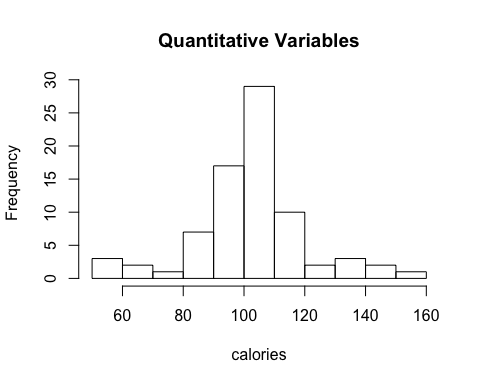


printcp(pruned.ct)

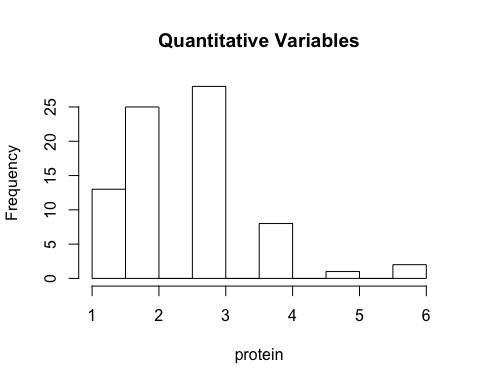
##   
## Regression tree:  
## rpart(formula = rating ~ ., data = Cereals2.df, method = "anova",   
## cp = 1e-05, minsplit = 5, xval = 5)  
##   
## Variables actually used in tree construction:  
## [1] calories cups fat fiber potass protein sodium sugars   
## [9] vitamins  
##   
## Root node error: 14997/77 = 194.76  
##   
## n= 77   
##   
## CP nsplit rel error xerror xstd  
## 1 0.5317148 0 1.000000 1.03624 0.211948  
## 2 0.1813001 1 0.468285 0.60397 0.139849  
## 3 0.0558501 2 0.286985 0.40080 0.088343  
## 4 0.0388397 3 0.231135 0.41989 0.088549  
## 5 0.0327546 4 0.192295 0.33777 0.076374  
## 6 0.0215222 5 0.159541 0.36524 0.092965  
## 7 0.0128738 6 0.138018 0.34436 0.091881  
## 8 0.0116026 7 0.125145 0.32977 0.079901  
## 9 0.0096056 8 0.113542 0.33224 0.079859  
## 10 0.0081724 9 0.103937 0.32564 0.080166  
## 11 0.0069790 10 0.095764 0.34158 0.083083  
## 12 0.0046084 11 0.088785 0.33130 0.083169  
## 13 0.0043413 12 0.084177 0.34175 0.083455  
## 14 0.0041531 14 0.075494 0.34309 0.084342  
## 15 0.0041072 15 0.071341 0.34309 0.084342  
## 16 0.0038255 16 0.067234 0.33811 0.084053  
## 17 0.0025201 17 0.063408 0.33232 0.075634  
## 18 0.0019030 18 0.060888 0.32577 0.075614  
## 19 0.0017860 19 0.058985 0.32596 0.075452  
## 20 0.0012123 20 0.057199 0.32071 0.075466

# Histogram for all quantitative variables

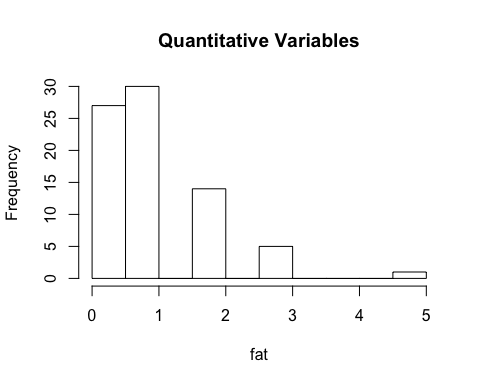
hist(Cereals2.df$calories, xlab = "calories", main = "Quantitative Variables")



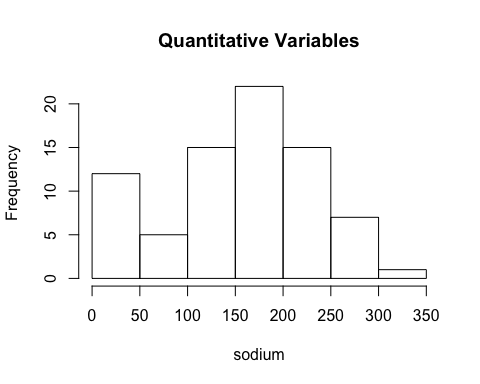
hist(Cereals2.df$protein, xlab = "protein", main = "Quantitative Variables")



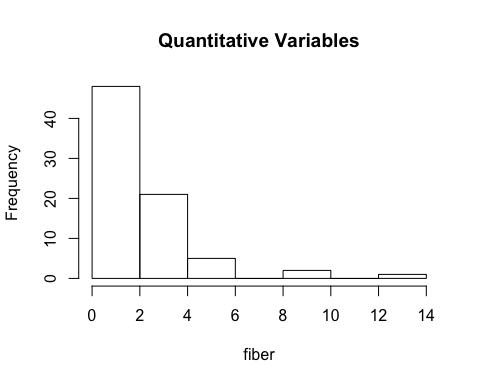
hist(Cereals2.df$fat, xlab = "fat", main = "Quantitative Variables")



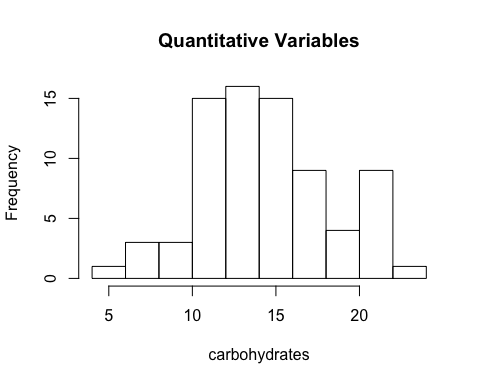
hist(Cereals2.df$sodium, xlab = "sodium", main = "Quantitative Variables")



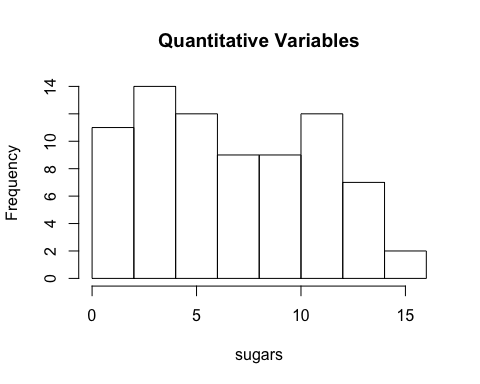
hist(Cereals2.df$fiber, xlab = "fiber", main = "Quantitative Variables")



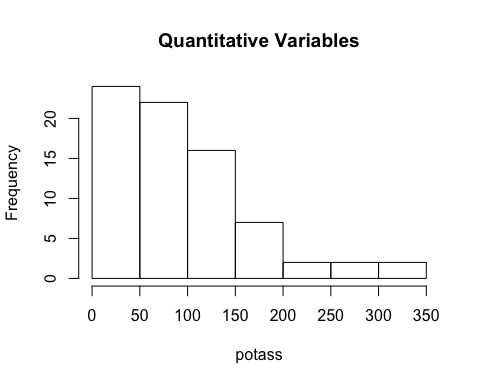
hist(Cereals2.df$carbo, xlab = "carbohydrates", main = "Quantitative Variables")



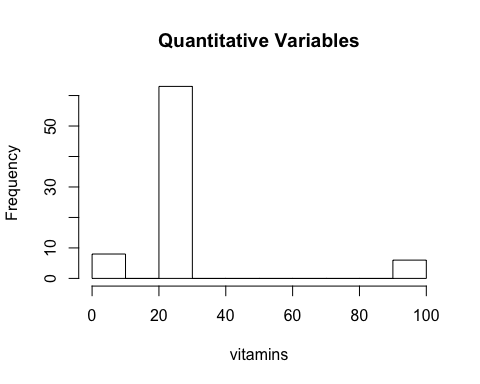
hist(Cereals2.df$sugars, xlab = "sugars", main = "Quantitative Variables")



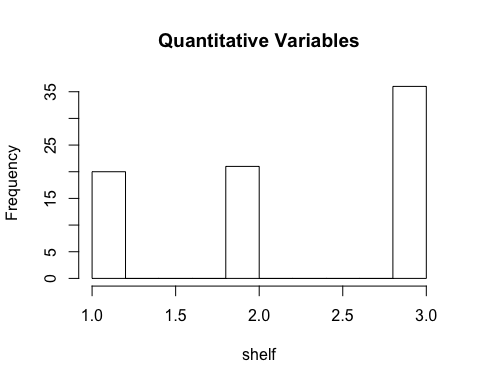
hist(Cereals2.df$potass, xlab = "potass", main = "Quantitative Variables")



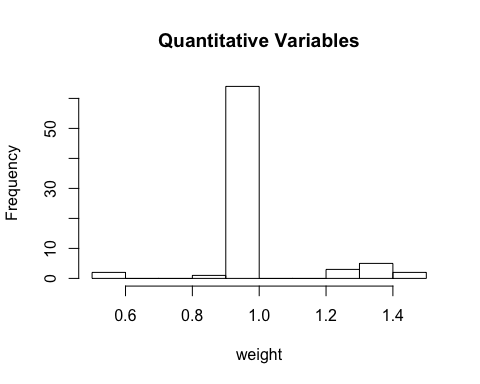
hist(Cereals2.df$vitamins, xlab = "vitamins", main = "Quantitative Variables")



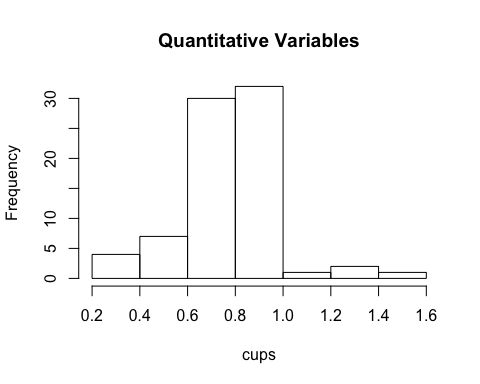
hist(Cereals2.df$shelf, xlab = "shelf", main = "Quantitative Variables")



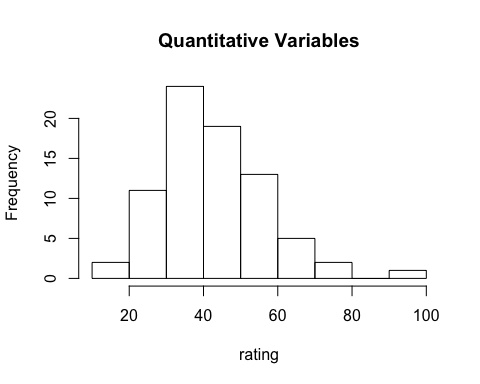
hist(Cereals2.df$weight, xlab = "weight", main = "Quantitative Variables")



hist(Cereals2.df$cups, xlab = "cups", main = "Quantitative Variables")



hist(Cereals2.df$rating, xlab = "rating", main = "Quantitative Variables")



# Boxplot for Customer satisfaction per shelf height

boxplot(Cereals2.df$rating ~ Cereals2.df$shelf, xlab = "Shelf Height", ylab = "Cust Rating",   
 main = "Customer Satisfaction \n and Shelf Height")

